Data Capture using FAX and Intelligent Character and Optical character recognition (ICR/OCR) in the Current Employment Statistics Survey (CES)

Richard Rosen, Vinod Kapani, Patrick Clancy

Key Words: Character Recognition; FAX; Automated processing

Purpose:
This paper reports on a feasibility study to determine whether Intelligent Character Recognition (ICR) or Optical Character Recognition (OCR) engines and related tools can be used to reduce/eliminate various aspects of survey processing currently associated with mail and FAX operations in the Current Employment Statistics (CES) program.

CES currently collects over 30,000 reports each month via FAX. These reports are collected on a form that is FAXed to the respondent each month. The respondent hand-enters the data and FAXes the form back to BLS. An operator retrieves the FAX and keys the data into a database. The computer systems performs a series of validation edits. In addition, BLS collects about 12,000 forms via mail. Here, forms are mailed to the respondent each month, filled out and mailed back. for data entry.

The study concludes that current ICR/OCR engines are sufficiently robust to meet most of the needs for the CES program and could provide significant cost savings while maintaining the accuracy of the data.

Background:
Several years ago, CES staff looked for ICR/OCR technologies for data collection, but at that time the recognition rate for the hand written numbers was not high enough to use in the production environment. Now, it appears the software has come of age, so we decided to take a fresh look into various available products that could recognize handwritten numbers with a high confidence. Several products were made available to our group for evaluation. The following paragraphs will try to elaborate on this evaluation and comments about the features, usefulness, cost benefit and their efficiencies in reference to labor and time saving. The ease of deployment of this software in the production environment is also evaluated.

CES Requirements:
CES has several unique requirements that significantly increase the difficulty of using character recognition technology. First, the characters are hand-written by respondents onto the CES form. CES collects 5 or 6 basic data elements from it's sample units; total employment, number of woman employees, number of production or nonsupervisory employees, and the payroll and hours for nonsupervisory employees. An average "line" of CES data contains about 25 characters (numbers) when a respondent is reports all five CES data elements. Second, the FAX to be processed in most instances has been FAXed twice. That is, we FAX the form to the respondent using a batch FAX system. The respondent then completes the form and FAXes it back to BLS. This double-FAXing reduces the quality of the image to be processed. Third, most respondents that report via FAX are multi-unit establishments; that is, they report for more than one location. On average firms that use the current multi-FAX form report for an average of 6 locations. In order to be efficient and not send multiple pages to the respondent, the FAX form must have space to enter at least 8 lines of data (one line for each location being reported).

The ICR/OCR Technology:

Before reviewing various products and recognition results, it is instructive to review the basic components of an ICR system. In order to have a complete ICR system, the following processes are necessary for successful recognition and data capture.

1) Template/Form creation:
Template or form creation is necessary for any recognition engine. This is necessary to predefine the forms for its identification (images received may have to be rotated, deskewed etc), position of data fields, type of fields, length of field, type of recognition and other properties associated with any data item which we want to recognize. After defining the template, the template is registered for

recognition. It appears that the current CES multi unit fax form can be tweaked for the purpose. Attachment 1 show the current multi unit fax form in use, while attachment 2 shows the new prototype fax form for use with ICR/OCR.

2) Image Capture & Scan:
The faxes are received on a FAX PC or on paper. If the fax is received on paper, then it must be scanned and store on the processing PC for recognition. If the fax is received directly on a FAX PC, then the image file can be copied manually or automatically to the processing PC. Capturing the image electronically on the PC provides for greater efficiency (no manual handling of paper and no need to scan the images

3) Image pre processing:
This process analyzes the captured image for the quality, form identification, screen displays etc. This is achieved by one of the following; image rotation, deskewing, despeckling, border removing, edge noise or entire background, locating data field and their identification and registration of the form with the existing form template. The form templates are created with predefined data fields and their locations on the form. The FORM Generator software can be used for the purpose. If the image identification fails, i.e. if the image does not match with the existing template, then the form is rejected and put in to the rejected folder for manual intervention. Our experiment suggests this is a critical step in the process, especially since the CES form contains many lines of data and goes through a double fax process, making the resulting image more difficult to recognize then other faxes.

4) Batch Processing:
Once the images are identified according to the preset templates, batches are created for recognition and passed to the recognition module. These batches can be created based on form type and /or the number of images in a batch. Generally it is most efficient to wait until sufficient quantity have been received, so the reviewer can be more productive.

5) Recognition:
The recognition engine is at the heart of any productive ICR/OCR system. The recognition engine starts processing one form at a time from the batch folder and marks the data fields in different colors depending on the recognition confidence. The recognition engines are capable of recognizing the free hand writing, hand written numbers and alphabets, check boxes, bar codes, OCR characters and magnetic ink characters. The recognized images then stored as images as well as proprietary files and passed to the data validation and review module for further processing.

6) Data Validation and Review:
This is a post recognition processing module. After the recognition is complete, the data verification operator then reviews the batch. In a split screen the actual image and the recognized data are shown. The operator can scan the image with the naked eye, check and correct any unrecognized data elements. Secondly, the operator can review the verified fields for accuracy as well as false positive type data fields. Once the batch validation is finished, the data can be exported to various data formats e.g. flat ASCII file, PDF, HTML or ODBC compliant data bases.

Technology behind the Recognition engines:
The recognition engines are based on Neural Network software technology, which converts the hand printed, machine generated character fonts, bar codes etc., into digital characters readable and usable by other computer application. With the help of pre processing of the fax or other image formats, the confidence levels of recognition has increased considerably. In our case, this technology will help us in recognizing the hand written numbers in the specified boxes on a form and convert them to ASCII text to minimize the data entry from the forms.

There are two types of errors, rejection and substitution. In rejection the system is unable to read the character and does not recognize the character at all. With substitution error, the system miss recognizes the character and substitutes the character with high level of confidence. This error may also be called “false positive”. Therefore, the main challenge is:
a). To obtain a high recognition rate (to reduce manual entry by the reviewer).
b). To obtain a low false positive rate (to avoid passing incorrect data to the analyst and to the estimation).

The false positive or the substitution errors can be minimized by using several techniques. First, by predefining the thresh holds or confidence levels for rejection when fields are defined. That is, you

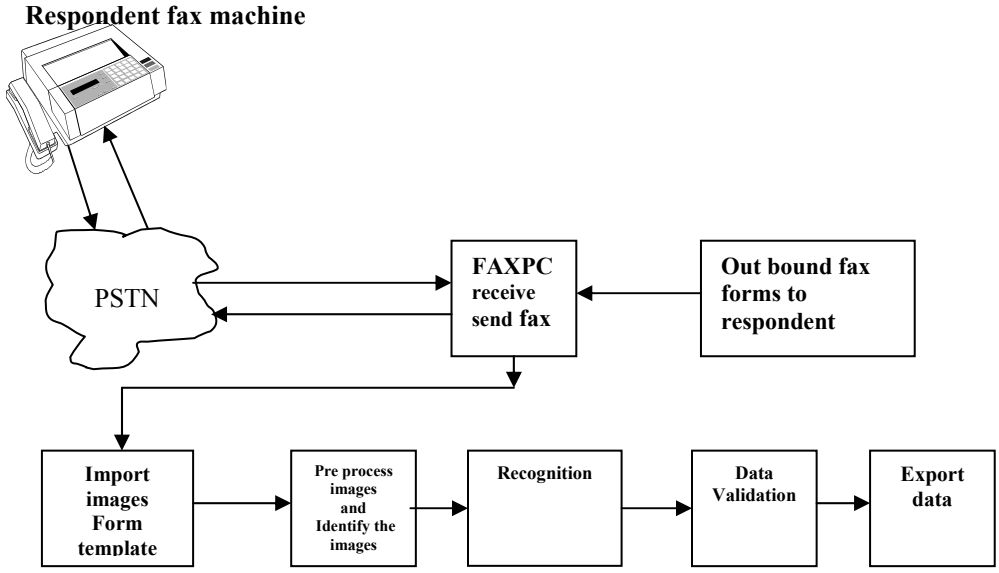
can set the confidence levels very high for recognition so that any character that is not recognized at, say, 95%, is marked for review. Another way to reduce false positives is to process the image twice for recognition; first with the raw image and then dilate the image and process again. These results of the two passes can be compared and any differences flagged for review in addition to any character that failed either scan/process. A third tool to minimize false positives is the application of logic edits during validation. For example, ensuring WW and PW are equal to or less then AE and AHE and AWHs are within acceptable limits. Records that fail these tests would be flagged even if the recognition engine passed them.

It should be noted that our current testing did not utilize such checks, and the false positive rate was under 5%.

The following diagram is the process flow of the Fax Forms sent to the respondents.

Figure 1:

ICR/OCR FORMS FLOW and PROCESSING DIAGRAM



Limitations:

We contacted a number of software vendors, who are working in this specialized field of intelligent character recognition for an evaluation package. Unfortunately, only two companies provided us with the package. Mitek allowed us to evaluate their full product as well as the tool kit. ParaScript provided us with their tool kit for creating the processes defined above. The Visionshape company markets Autoscan32 software. The demo was provided for us to evaluate, this product uses the fine Reader engine by ABBEY, Mitek uses Quickstrokes and Pegasus uses Smart Scan engine for ICR/OCR. Another company which claims to have a full solution with all the processes (CAPTIVA SOFTWARE), was unable to provide to us either demo or evaluation package. We did not get any reply from the company representative after several tries.

The following table depicts the limited comparison of different processes, the cost of development and the time to create all the needed functions using the company's software package, tool kits and original programming.

Table: 1

Features/Company	**Mitek**	**Mitek**	**Pegasus**	**Vision shape**	**ParaScript**	**Abbey**
Product	**Doctus+API**	**QuickStrokes API**	**SmartScan**	**VisionTools**	**FieldScript**	**Finereader API**
Icr/Ocr engine	QuickStrokes	QuickStrokes	SmartScan	Fine Reader	FieldScript	Fine Reader
Fax Image Normalization	Yes	Yes	3rd Party	3rd Party	3rd Party	3rd Party
Field Definition	Yes	Yes	CODE	VisionTools	CODE	CODE
Form registration	Yes	Yes	CODE	VisionTools	CODE	CODE
Batch Creation	Yes	CODE	CODE	VisionTools	CODE	CODE
Field recognition[1]	Yes	Yes	SmartScan	VisionTools	CODE	CODE
Data Validation[2]	Yes[3]	CODE	NA	VisionTools	CODE	CODE
Data export	Text,Oracle DB	CODE	CODE	TEXT,ODBC	TEXT,ODBC	CODE
Accuracy(1-10)[4]	9	9	8	5	7	5
Ease of Integration(1-10)	7	9	8	6	7	7
Time to customize	1~2 months	5~6 months	7 months	8 months	6~8 months	7~8 months
Cost of product	46,000+6912/yr[!]	26,200+3300/yr[!]	3000+775*	7000+900**	No quotes	No quotes

[1]Refers to ICR performed on pre-defined field
[2]Refers to Data validation and correction performed by Operator
[3]Requires additional customization
[4]Relative observed accuracy of ICR
[5]Relative ease of integration into existing workflow
[!]Maintenance per year
[*]Additional run-time license per processor
[**]Support fee
CODE We have to write code for this

Mitek's Doctus package (which includes the API's needed for full customization) would provide both the highest recognition and shortest development time. Mitek's QuickStrokes API package is less costly alternative. It provides the same recognition capabilities, and the ability to fully customize the system. However the development time would be a bit longer as more modules would need to be programmed in–house.

ICR Evaluation:
We used the following method to compare the ICR process with the current manual data entry system.

Methodology:
A sample of 147 reporters was selected randomly by obtaining 15 filled fax forms from the Atlanta Data Collection Center, which were received earlier and manual data entry done on the forms.

The data on these forms was re-written by different MIES staff on modified fax forms. These forms then faxed to an NT PC server to capture the image for processing. Once the faxes were received, the required ICR processes were performed sequentially. The results were then tabulated to compare the accuracy of the ICR against the actual data. Two tests were performed on the received faxes. The first test was setting the global parameter of dilation so that the character's thickness is enhanced from the raw image and second test was without dilation. The second test did not use dilation. We found that dilation provided enhanced recognition.

Results: The following graphs will show the recognition/non recognition of the characters per record in these tests.

Figure 1:

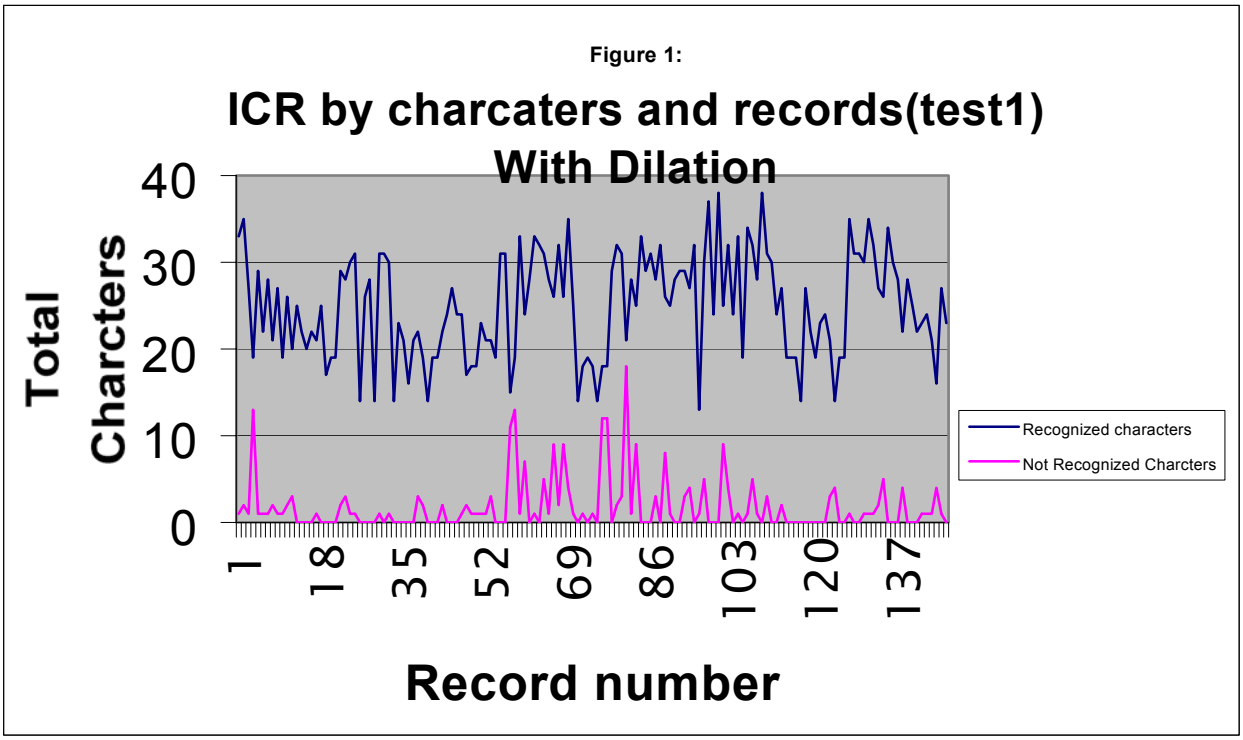


Table 2:

The table below shows the average characters per record and recognition.

Average character per record 26.65	**Recognized**	**%**	**Not Recognized**	**%**
Test1 (dilation)	24.88	93.38	1.77	6.62
Test2 (no dilation)	23.31	87.31	3.39	12.69

The following tables depict the false positive % or substitution error % value in both the tests.

Table 3: Test 1 (dilation)

value/score	>=900	<900	Total
actual=recog	515	**307**	822
Actual Not Recog	**32**	86	118
	547	393	940
False+%	3.4		

Table 4. Test 2 (no dilation)

value/score	>=900	<900	Total
actual=recog	445	298	743
Actual Not Recog	94	103	197
	539	401	940
False+%	10.00		

Note: 900 refers to the internal confidence level set to determine recognition. For example, in Test 1, the system passed/recognized 547 characters. Of those, 32 were actually incorrect, for a false positive rate of 3.4%

Table 3 clearly demonstrates that with appropriate field definitions and setting of recognition parameters the false positive cases can be reduced to the minimum level. They may completely be removed by adding field validation checks for the ValEdit module (process) for the operator. What is also interesting here is most false negative (307) fields are actually recognized correctly.

The table 5 compares the time required for manual review/validation of ICR against average current data entry time in SOLCATI. This provides a

benchmark to evaluate the relative efficiency of ICR/OCR review time with current data entry times.

Table 5: Data validation and data entry using the ICR package and actual data entry in SOLCATI			
Mode data entry/validation	Total processing time in Minutes For the 14 sample forms.	Secs/record	Manual data entry time in secs/ record/batch
Trial run with eye scan	21	8.57	95.5 secs/ rec/ 1 record 45.0 secs/ rec/ 4 records 34.5 secs/ rec/ 8 records
Actual run with eye scan	15	6.12	
Invalid Fields, Partial scan	8	3.27	
Invalid fields only	6	2.45	

The ratio of reviewing 8 records of ICR data with actual run and eye scan of the image versus current manual data entry times with a trial run and eye scan of the image is 1:3. With partial scan and invalid fields only is 1:8.5 and for invalid fields only 1:11. If the operator used ICR package and performed data validation/entry on invalid fields only, it is 11 times faster then the manual entry.

These comparisons suggest considerable savings in staff time for data entry with ICR.

CES currently receives approximately 18,000 fax reports. Table 6 summarizes the details of a cost comparison for labor for data entry using ICR and the current SOLCATI system.

Table 6: Comparison of costs of manual and ICR data entry				
			Manual Data entry (hrs.)	ICR data entry (hrs.)
Total Records to process	18,000		210	45
Cost per hour	$23	Total Cost:	$4,830	$1,035
Total forms to process	2,250			
Peak day-Forms received	375			
No. of records on peak day	3,000			
Time required for data entry(hrs)	8			
Total Saving per month	$3,795			
Total Saving per year	$45,540			

Conclusion:

From the above review we concluded that ICR technology along with our FAX system is a viable cost-effective option for processing CES reports presently being received via paper FAX. Overall, we estimate that the cost of the ICR system would be recovered with in a year. Fax ICR technology is sufficiently accurate to provide high quality data with much less operator time than is presently used for full key entry of the data.

Attachments:

The following two pages compared the current multi-FAX form in use with a slightly modified form that is more OCR friendly. Several minor changes were made. First, the number of lines of data per page was reduced from 12 to 8. This allowed for more white space to differentiate the characters during recognition. Second, the "location" description was combined with the report number box. This provided additional horizontal space for entry of digits. This was also needed to improve recognition. Third, "registration marks" (the large black dots) were added so that the image could be rotated and properly aligned. Fourth, a different, more OCR readable, font was substituted for the report ID.

Note: the forms as shown are not the actual size of the forms. The forms print on standard 8.5"x11" paper.

Bureau of Labor Statistics: Report on Current Employment Statistics – **FAX Report Form** **U.S. Department of Labor**

Form Approve
O.M.B. No. 1220-001

Firm AT Contact: Mr. James Mattson Title: 54880

Telephone number:2 Mass Ave. N.E. #4860 Fax Number: Washington

This report's reference month and year:

Column 11 indicates how often your Production, Construction, or Nonsupervisory Employees are paid:	**1** Each week	**3** Twice a month	**5** Other
	2 Every two weeks	4 Once a month	

Please fax this report to: 405 W 1st St by

Please complete columns 4 - 10 only for the pay period which includes **the 12th of the month**. Complete column 9 for establishments in *trade* industry only and column 10 for establishments in *manufacturing* industry only. Please enter a comment to explain significant changes in your employment in column 12. Detailed definitions and instructions are on the previous two pages of this form. Please round all data to the nearest whole number (**omit cents and fractions**). After completion, please fax this form to the FAX number shown above. Your participation is very much appreciated.

CES Report Number (1)	State (2)	Establishment Location (3)	All Employees (4)	Women Employees (5)	Production, Construction, or Nonsupervisory Employees (6)	Payroll ($) (omit cents) (7)	Hours (omit fractions) (8)	Nonsupervisory Employee Commissions ($) (trade only) (9)	Production Worker Overtime Hours (manufacturing only) (10)	Length of Pay Period (11)	Comment Code (see definitions) (12)
2097349		715-392-1755						0		1-800-	

Did you open or close an establishment? Need help? If you open/close an establishment, or the pre-printed information on this form is incorrect, or you need help filling out this form, please contact us at.

PC

BLS-790 FI FAX Report Form Jan 98

Attachment 1:

Bureau of Labor Statistics: Report on Current Employment Statistics – **FAX Report Form Page 1 of 2**

U.S. Department of Labor

Firm:	TEST FIRM 24	Contact:	Ms. Laura Jackson	Title:	Payroll Manage	FAXFORM TEST V. 5	Form Approved O.M.B. No. 1220-0011
Telephone number:404-331-1950		Fax Number:	202-691-6566		**This report's reference month and year:**		10/2001

Column 11 indicates how often your Production, Construction, or Nonsupervisory Employees are paid: 1 Each week; 2 Every two weeks; 3 Twice a month; 4 Once a month; 5 Other

Please fax this report to: 1-800-239- **by** 10/26/2001

Please complete columns 4 - 10 only for the pay period which includes **the 12th of the month**. Complete column 9 for establishments in *trade* industry only and column 10 for establish-ments in *man-ufacturing* industry only. Please enter a comment to explain significant changes in your employment in column 12. Detailed definitions and instructions are on the previous two pages of this form. Please round all data to the nearest whole number (**omit cents and fractions**). After completion, please fax this form to the FAX number shown above. Your participation is very much appreciated.

CES Report Number State Establishment Location (1)	All Employees (4)	Women Employees (5)	Production, Construction, or Nonsupervisory: Employees (6)	Payroll ($) *(omit cents)* (7)	Hours *(omit fractions)* (8)	Nonsupervisory Employee Commissions ($) *(trade only)* (9)	Production Worker Overtime Hours *(manufacturing only)* (10)	Length of Pay Period (11)	Comment Code *(see definitions)* (12)
860074211 NY Mining Stone-Orisk Fls						TTTTT	TTTTT		
860074212 NY Statewide Const Gen L Office-Oneida						TTTTT	TTTTT		
860087066 NY Mining Stone-Jordanville						TTTTT	TTTTT		
860087067 NY Sand & Gravel-Poland						TTTTT	TTTTT		
860087068 NY Black Top Plant-St Johnsville						TTTTT	TTTTT		
860087069 NY Black Top Plant-Whitesboro						TTTTT			
860087070 NY Black Top Plant-Herkimer County						TTTTT			
860087072 NY Unit 11-Auburn						TTTTT			

Need help? If you open/close an establishment, or the pre-printed information on this form is incorrect, or you need help filling out this form, please contact us t1-800-347-3764. PC 16

BLS-790 F1 FAX Report Form Jan 98

FAXFORM 2

Attachment 2: